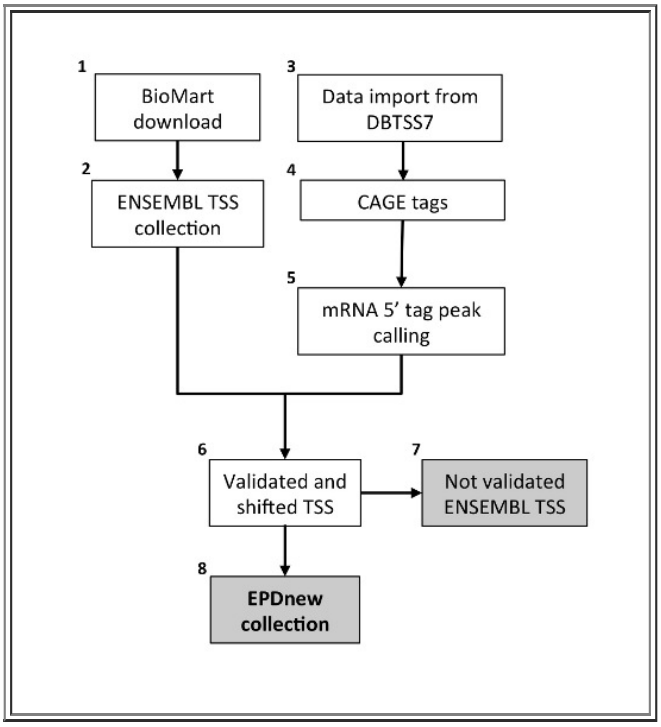
TSS assembly pipeline for Mm_EPDnew_001

Introduction

This document provides a technical description of the transcription start site assembly pipeline that was used to generate EPDnew version 001 for *Mus musculus* genome assembly mm9.

Source Data

Assembly pipeline overview



Description of procedures and intermediate data files

1. Biomart Download

Data was downloaded from dec2011.archive.ensembl.org/biomart/martview/ selecting the following attributes:

1. Ensembl Gene ID
2. Ensembl Transcript ID
3. Chromosome Name
4. Strand

5. Transcript Start (bp)
6. Transcript End (bp)
7. Gene Start (bp)
8. Gene End (bp)
9. Status (transcript)
10. Status (gene)
11. Associated Gene Name

Then, transcrips have been filtered according to the following rules:

1. Transcripts of protein coding genes only
2. Transcript length > 0 [Transcript Start different from Transcript End]
3. Transcript lies on full chromosomes
4. Gene must have a 5' UTR [Transcript Start different from Gene Start]
5. Genes must be annotated [Associated Gene Name present]
6. Gene and transcripts status known

Gene names were taken from the field "Associated Gene Name". Since the EPD format doesn't allow gene names longer than 18 characters, we checked whether the names repsected this limitation.
Transcripts with the same TSS position were merged under a common ID. As a conseguence of this, from the 95884 transcrips originally present in the ENSEMBL database, ~30000 were merged, leaving 67440 uniquely mapped promoters in the input list.

2. EMBL TSS collection

The ENSEMBL TSS collection is stored as a tab-deliminated text file conforming to the SGA format under the name:

Mm_ENSEMBL65.sga

The six field contain the following kinds of information:

- NCBI/RefSeq chromosome id
- "TSS"
- position
- strand ("+" or "-")
- "1"

- gene name.

Note that the second and forth fields are invariant.

3. Data import from DBTSS7

Solexa Tag Data were downloaded from DBTSS ftp-site (see link above). The source files are the following:

- 3t3_data.tab.gz: Mouse 3T3 Solexa tag mapping data;

According to the readme file included in the ftp archive, the 5' end tags were mapped to the Human genome hg18. The source format is a non-standard tab-delimited format that has been converted to SGA via an ad hoc perl script. All tissues have been merged into a single file.

4. CAGE tags

The compressed version of this file is available from the MGA archive (see above) under the name:

dbtss7.sga.gz.

5. mRNA 5' tags peak calling

Peak calling for the merged file has been carried out using ChIP-Peak on-line tool with the following parameters:

- Window width = 100
- Vicinity range = 200
- Peak refine = Y
- Count cutoff = 9999999
- Threshold = 10

The sga file containing the list of peaks can be downloaded here.

6. TSS validation and shifting

The source data (mRNA peaks and ENSEMBL TSS) was then processed in a pipeline aiming at validating transcription start sites with mRNA peaks. An ENSEMBL TSS

was experimentally confirmed if an mRNA peak lied in a window of 100 bp around it. The validated TSS was then shifted to the nearest base with the higher tag density. After this step, the total number of validated promoters was 9773.

The list of validated and shifted promoters can be downloaded here.

7. ENSEMBL not-validated TSS

The total number of non mRNA validated TSS was 57667.

8. EPDnew collection

The 9773 experimentally validated promoter were stored in the EPDnew database that can be downloaded from our ftp site. Scientist are wellcome to use our other tools ChIP-Seq (for correlation analysis) and SSA (for motifs analysis around promoters) to analyse EPDnew database.